MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo 334 June 1975

ANALYZING NATURAL IMAGES

a computational theory of texture vision

by

D. Marr

ABSTRACT: A theory of early and intermediate visual information processing is given, which extends to about the level of figure-ground separation. Its core is a computational theory of texture vision. Evidence obtained from perceptual and from computational experiments is adduced in its support. A consequence of the theory is that high-level knowledge about the world influences visual processing later and in a different way from that currently practiced in machine vision.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643.

Summary

Understanding how the visual cortex analyzes natural images is one goal of visual neurophysiology. At some stage, we need to confront the information processing problems that are involved. A series of computational experiments on natural images was therefore undertaken, and a visual pre-processor emerged with the following structure:

(1) Approximations to the first and second directional derivatives of intensity are measured everywhere. They are computed by convolving the image with "edge-shaped" and "bar-shaped" masks.
(2) These measurements are parsed into an orientation-dependent description of the intensity changes present in the image. The parsing process consists of discovering and matching peaks and troughs in the measurements, and roughly classifying local patterns of peaks into EDGEs, LINEs, SHADING, etc.
(3) The descriptions obtained at each orientation are combined, termination points of edges are discovered, and small blobs are isolated and described.

This pre-processor computes what is called the primal sketch of an image, but for most images it is large and unwieldy. By examining our ability to interpret certain simple drawings, it is demonstrated that a variety of abstract grouping processes and related facilities are present in our visual systems. It is shown how, if applied to the primal sketch, these processes are capable of successfully analyzing many kinds of visual texture, and of extracting perceived "figure" from ground. It is conjectured that these operations can account for the entire range of texture discriminations of which we are capable, and the analysis of several real images is given in its support. The conjecture relegates the influence of higher-level knowledge on visual processing to a much later stage than is currently found in machine vision programs, and it implies that such knowledge should influence the control of, rather than the actual computations in, the earlier stages of analysis.

Preface

The work of Barlow (1953), of Mountcastle (1957), of Lettvin et al. (1959), and of Hubel and Wiesel (1962) initiated what is widely regarded as a breakthrough in visual neurophysiology. But despite the subsequent accumulation of a wealth of anatomical and physiological information about the mammalian visual cortex, our knowledge of its information processing function, or even of how difficult are the problems that it solves, remains rudimentary.

This is no accident. Physiology has always been concerned with how organisms work. Its goals are to unravel the local mechanisms within an organism and to understand their place in the functioning of the animal as a whole. While the concerns of physiology lay with mechanical, or even with chemical or physical phenomena, the physiologist's background knowledge and everyday experience sufficed to provide him with the necessary insight into function. As physiology has turned to information processing problems, however, neurophysiologists have lost the reliable background intuition that has been fundamental to the success of the discipline in the past. The situation in modern neurophysiology is that people are trying to understand how a particular mechanism performs a computation that they cannot even formulate, let alone provide a crisp summary of ways of doing. To rectify the situation, we need to invest considerable effort in studying the computational background to questions that can be approached in neurophysiological experiments.

Therefore, although the work described here arises from a deep commitment to the goals of neurophysiology, the work is not about neurophysiology directly, nor is it about simulating neurophysiological mechanisms: it is about studying vision. It amounts to a series of computational experiments, inspired in part by some findings in visual neurophysiology. The need for them arisee because, until one tries to process an image or to make an artificial arm thread a needle, one has little idea of the problems that really arise in trying to do these things. Computational experiments allow one to study in detail what combination of factors causes a method, or group of methods, to succeed or fail in a number of particular circumstances that originate from real-world data. The power of this approach is that the knowledge one obtains concerns facts that are inherent in the task, not in the structural details of the mechanism performing it. Such knowledge is a vital prerequisite for understanding mammalian visual systems fully, and it is knowledge that cannot be obtained in any other way.

Introduction

The vision problem begins with a large gray-level intensity array, and culminates in a description that depends on that array, and on the purpose for which it is being viewed. The question of interest is what has to go on in between. In this article, we shall restrict our attention to single frame, monochromatic, monocular images without specularities, reflections, translucency, transparency, or light sources; and we shall study some of the problems that arise in understanding early and intermediate levels of visual information processing.

Perhaps the best way of introducing the topic is to pose some questions:

(1) What is early visual processing for?
(2) How much of visual information processing can proceed using purely data-driven techniques?
(3) At what level and by what mechanisms may texture vision and figure-ground phenomena be implemented?
(4) When does higher level knowledge about the world have to begin interacting with purely data-driven processes?
(5) When and how does purpose have to influence what computations are made on an image?

Recent work in computer vision has tried to involve high-level knowledge about the world at a very early stage in the processing (Shirai 1974, Freuder 1975). The main motivations for this have been that it has proved very difficult to extract object boundaries from intensity arrays, and that strategic deployment of high-level knowledge about a scene can sometimes greatly reduce the computational effort required for primary image processing. This article opposes this trend, and makes three main arguments. The first argument consists of a demonstration that a very great deal of information may in fact be extracted from an image using knowledge-free techniques. The price one pays for this is prodigious computing power, and it involves programs that are considerably more complex than feature-point detecting routines. There can, however, be little doubt that our own visual systems do in fact possess enormous power (Thomas and Binford 1974, p 16). The second argument is that deciding what a low-level visual processor can and cannot deliver is a pre-requisite for useful research into "higher-level" problems of recognition. For example, the problem of recognizing and interpreting a scene has a very different flavor in vision systems with rich and with poor pre-processing abilities. The difference is almost as extreme as trying to make sense out of an English sentence with and without the benefit of a knowledge of English syntax. Hence, unless one has a firm idea about what pre-

processing is possible, one is in danger of expending effort on problems that, in a real sense, are not problems at all. The third argument is that our own perceptual apparatus probably contains a rich pre-processing ability. Hence if machine vision intends to say anything useful about those computations, it had better examine the lower problems first, and study the later ones when the peripheral processing has been solved. Otherwise one is conducting research without the benefit of data on which to test one's conclusions. This amounts to a reckless abandoning of precisely the new experimental tools that computer technology has made available, namely the ability to decide whether a computational theory successfully addresses the problems that arise in real-world data.

This article presents a theory of visual processing for its chosen class of images up to about the level of the figure-ground problem. Its main focus is a new computational theory of texture vision. The article gives a sufficient number of examples of processed images to establish that the theory is not obviously inadequate. The detailed and lengthy arguments that make a positive case for adequacy will appear elsewhere (Marr 1976). The argument is quite protracted, and relies on several main steps. Its overall thrust is that the first step of consequence in visual information processing is to compute a primal description of the image, and that all subsequent computations are implemented as manipulations of that description. In order that the reader may follow with ease the stages in the argument, I summarize the main steps here:

(1) The function of early visual processing is to compute a description of the gray-level changes present in an image in terms of a vocabulary of gray-level change primitives. These primitives consist of straight contour segments of various kinds (SHADING-EDGE, EXTENDED-EDGE, etc.), LINEs, BLOBs, and of various parameters bound to them such as FUZZINESS, CONTRAST or LIGHTNESS, POSITION, ORIENTATION, simple measures of their SIZE, and a specification of their TERMINATION points. This primitive description is obtained from the intensity array by knowledge-free techniques, and it is called the PRIMAL SKETCH. It differs from an array of feature points in a subtle way, which is explained in the text.

(2) From our ability to interpret drawings, one may infer the presence in our perceptual equipment of symbolic processes that are capable of grouping lines, points, and blobs together in various ways. Non-symbolic techniques, like examining the power spectrum of the spatial Fourier transform of the drawings, cannot account for these grouping phenomena, since the groupings are performed by mechanisms of construction rather than mechanisms of detection.

(3) For most images, the primal sketch is large and unwieldy. It can however be capably analyzed by a mechanism that has available the symbolic processes

discovered in step (2), together with the ability to select items out of the primal sketch on the basis of first-order discriminations acting on the principal parameters. Hence, it is argued, texture vision rests on grouping operations and first-order discriminations operating on the primal sketch, rather than on second order operations operating on the intensity array as suggested by Julesz (1975). It is further argued that the set of processes whose existence is necessary in order to explain our ability to interpret drawings, is also sufficient, when applied to the primal sketch, to explain the range of texture vision that is present in humans. Fourier and power-spectrum techniques on their own are certainly deficient, and probably also unnecessary.

(4) The extraction of a form from the primal sketch using these techniques amounts to the figure-ground computation. Except in difficult cases, this extraction can proceed successfully without calling upon higher level knowledge, and it precedes the computation of the shape of the extracted form. This has two important consequences. Firstly, the isolation and delivery of a form to subsequent processes does not depend on being able to assign an accurate high-level description to it; and secondly, because of this it is easy to compute rough descriptions of complex forms. This is probably essential for the fluency of subsequent analysis of shape.

(5) The extent to which higher level knowledge and purpose influences the processing up to this stage is very limited. There is at present no reason to believe that higher level knowledge is needed to compute the primal sketch at all; and its role in the extraction of form from the primal sketch can often be limited to deciding which form should be extracted. It is conjectured that in all cases, higher-level knowledge need be only weakly coupled to the processes that separated figure and ground. This relegates the use of higher level knowledge to a much later stage than is found in current machine vision programs, and simultaneously confines much of its impact to influencing control, rather than interfering with the actual data-processing that is taking place lower down.

Each step in the argument is treated in a separate section.

Early Processing: computing the primal sketch

The primal sketch consists of a primitive but rich description of the intensity changes that are present in an image. This description consists of a set of assertions, expressed in terms of a vocabulary of symbols and modifiers that are powerful enough to capture all of the important information in an intensity array. An example of such an assertion might be:

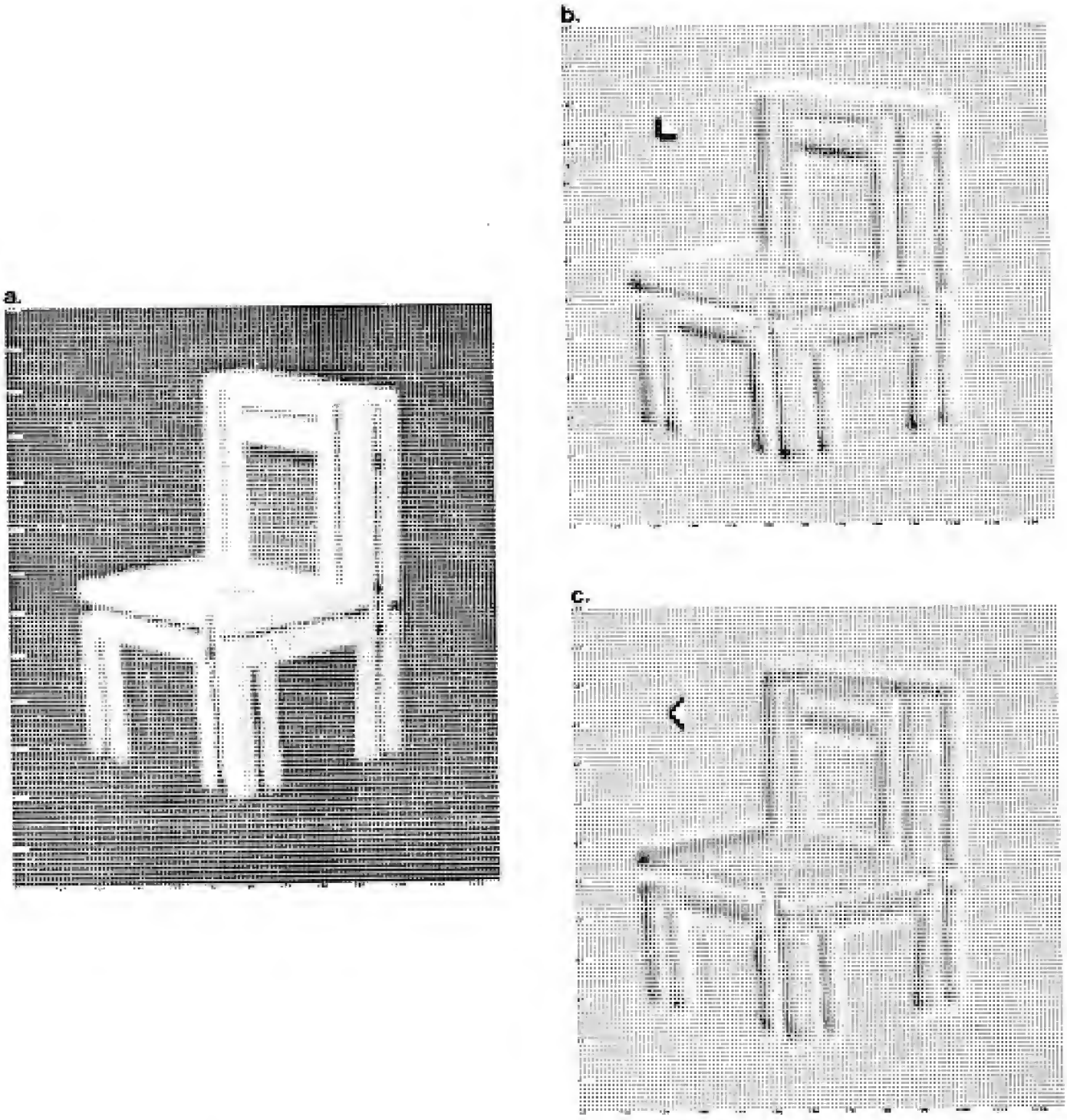
(SHADING-EDGE (POSITION (34 48) (73 48))
(CONTRAST 34)
(FUZZINESS 17)
(ORIENTATION 0))

The first problem is how such an assertion may be computed -- what measurements should one first make on an image, and how should those measurements be combined to enable the assertion to be made.

To help us answer these questions, let us see what neurophysiology tells us. Simple cells in the cat make measurements upon an image, and the nature of the measurement that they make is fairly well understood. Their receptive fields are either bar- or edge-shaped (Hubel and Wiesel 1962), and if other parameters are held constant, they signal the linear convolution of a bar- or edge-shaped mask with the intensity distribution currently falling upon the retina, in logarithmic units of contrast (Maffei and Fiorentini 1973, figure 8). Not all of what are now called simple cells behave linearly, but a distinct subclass does. The important question for understanding the analysis of visual information is whether these cells represent assertions other than the fact of the measurement itself; and if they do, what are they? One idea is, for example, that a cell with a bar-shaped receptive field signals an assertion about the presence of a bar in the visual field; but a moment's thought reveals that this is impossible, since such cells respond also to the presence of a single edge. Another puzzle concerns the existence of both bar-shaped and edge-shaped receptive fields (in different cells). Since both kinds detect changes in intensity, why are both types needed? The reason is probably that changes in intensity are not the only important types of change in an image -- changes in intensity gradient often provide important, and sometimes the only information that an object boundary is present (Marr 1974b). An edge that consists of a step change in intensity gradient rather than in intensity may be produced by a lambertian white cube aligned at 45 degrees to the viewer and illuminated from the viewing position. Perceptual evidence of our sensitivity to such edges is easy to find: Mach Bands are the most well-known example (see e.g., Ratliff 1965). This immediately suggests that one should regard simple cells that have an edge-shaped receptive field as measuring something like the first directional derivative of intensity; and those with a bar-shaped receptive field as measuring the second directional derivative. Two questions then arise: firstly, why compute directional measures? And secondly, what should one do with the measurements when one has them?

The application of a bar-shaped mask to an image does not, as we have seen, lead directly to an assertion about the presence of a bar in the image. The underlying point concerns the relation between computing the bar assertion, and the inverse transform of the original measurement, and it is a point of some importance. Let us consider the computation of an assertion about the presence of

FIGURE 1



1. The image of a chair (1a) has been convolved with two "corner-masks" (1b and 1c). The mask shapes are shown in the figures. Detecting corners from such measurements is not straightforward.

a corner in the image of figure 1a. A way of computing this assertion that immediately springs to mind is to take a specially "tuned" corner-shaped mask. One might conjecture that a "corner" exists in the image at a point P provided that the mask gives a value there which is greater than some threshold. Figures 1b and c show the convolution of corner masks with the image; but can the reader confidently distinguish the corners from these measurements? The reason for the failure is that the inverse transform to that produced by a corner-shaped receptive field depends critically on the boundary conditions that obtain. Any method that computes a corner assertion is saying something about this inverse, and so must take enough information into account at each point to satisfy the dependence on boundary conditions. This extra information may be provided by looking at the results of the corner-mask at neighboring points, or by looking at the results of some other measurement taken in parallel; the important point is that the computation is not a trivial one, and has to take these extra factors into account. It is not impossible to use primary measurements that are not orientation sensitive, but the extra computation involved is expensive, since one switches from having to look in just two directions to having to look in all directions. A persuasive case would have to be made if one were to choose a primary measurement that was not directionally selective.

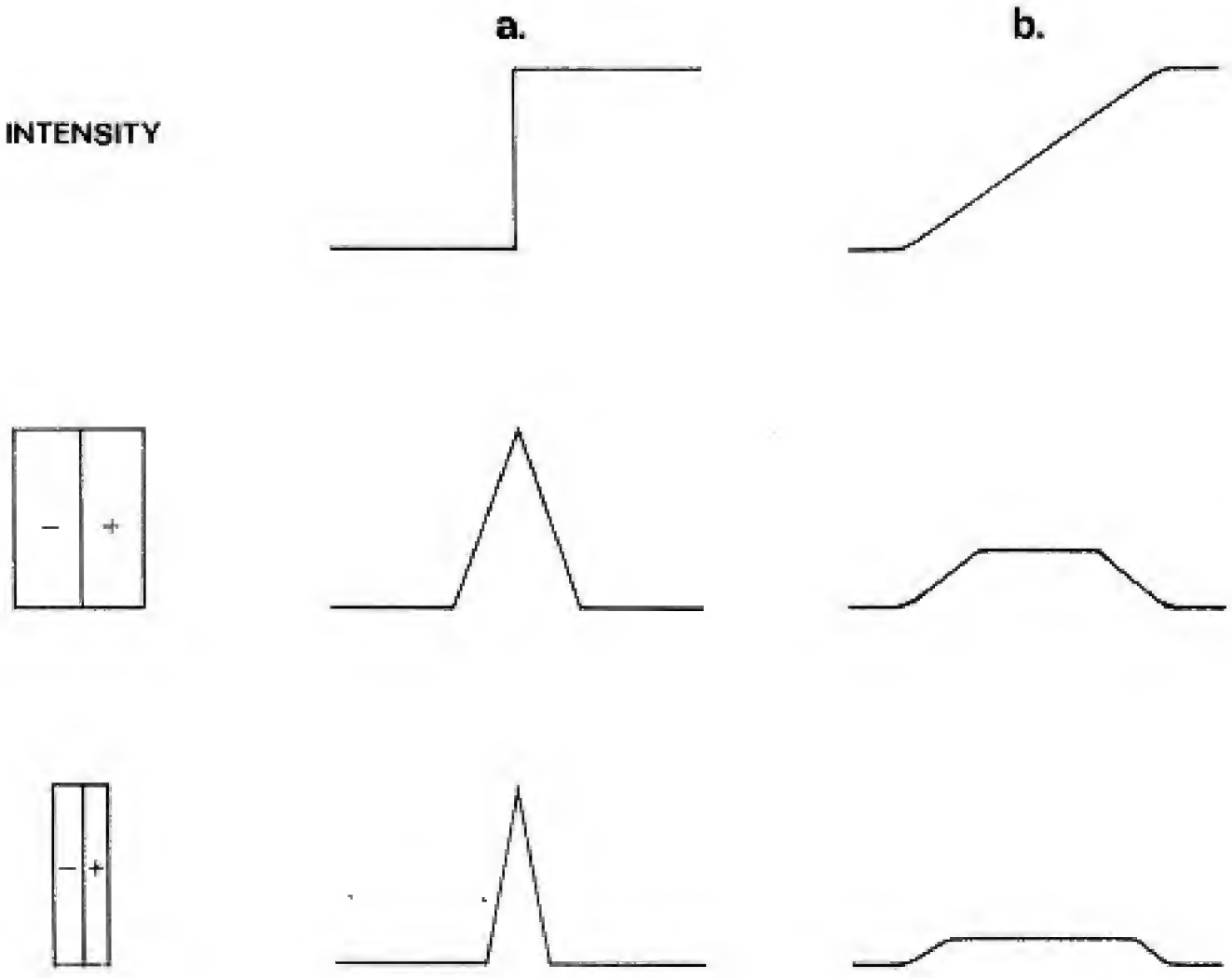
Translating the measurements into a description

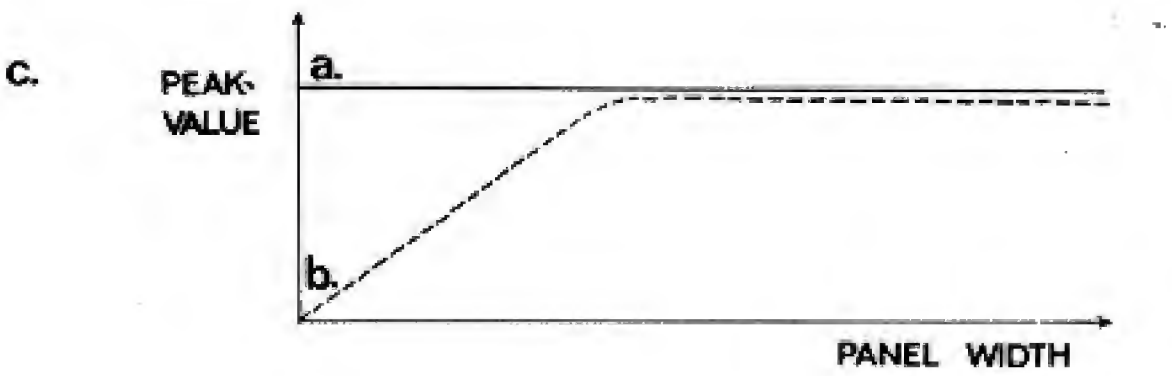
Suppose then that one measures the first and second directional derivatives of intensity everywhere in an image. What do we do with them? Translating one large array of numbers into several other large arrays is not an obviously useful process. It turns out, however, that we can make a great simplification at this stage in the analysis. Provided that measurements are made with masks of several sizes, one can show that the positions and sizes of the peaks in the measurements provide enough information to compute the description of the underlying intensity changes. Furthermore, provided that a group of peaks is sufficiently isolated from other peaks, the other peaks may be ignored when analyzing that group.

The reason for this is illustrated in figure 2, which shows the difference between edge-mask values obtained using masks of two different sizes on a step change in intensity (2a), and on a gradual change (2b). The results are analogous to the power spectra of different kinds of edge. Step changes are "seen" equally well by all sizes of mask. Gradual changes are seen increasingly faintly by edge-shaped masks whose dimensions are smaller than the distance over which the intensity change is taking place. Figure 2c shows this effect in graphic form, and from it one can see that a good estimate of the "fuzziness" of an edge may be made by finding the mask size at which the edge-mask response starts to

FIGURE 2

2. Diagrams of "edge-shaped" mask convolutions with a step (a) and with a gradual (b) intensity change. The intensity profiles appear at the top. The convolutions with the two sizes of mask shown on the left appear beneath the intensity profiles. For a step change in intensity, masks of all sizes produce the same maximum response (trace a in graph (c)). Gradual intensity changes are seen progressively weaker by the smaller masks (trace b in graph (c)).


a.
b.
INTENSITY
−
+
−
+


c.
PEAK-
VALUE
a.
b.
PANEL WIDTH

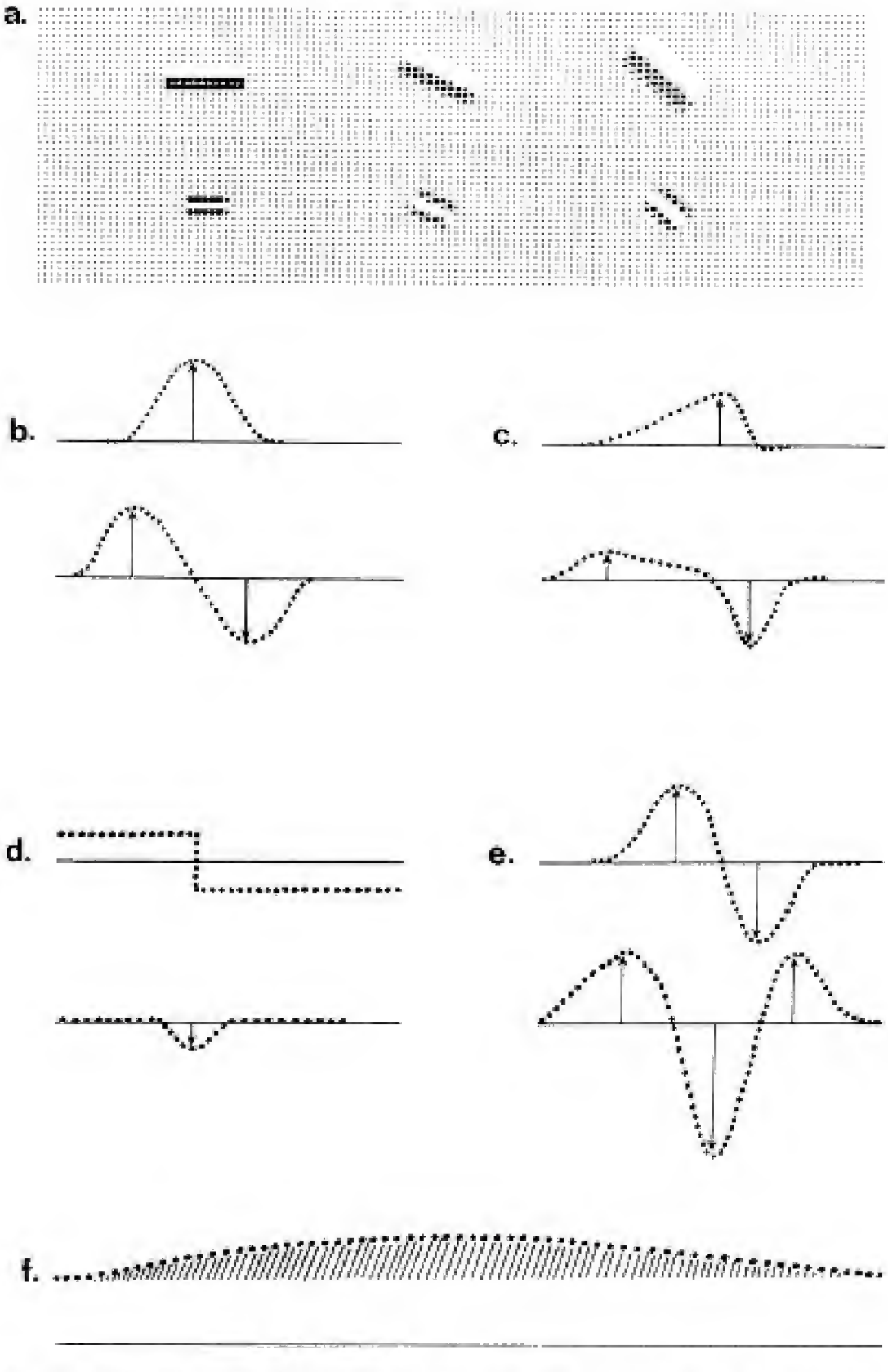
diminish.

This shows one way in which the use of multiple mask sizes is important, but there is another reason which is perhaps even more important. It is that where a faint edge exists in the image, it is frequently impossible to tell from a single record which of the peaks are important, and which are due to noise. Matching peaks obtained using different sizes of mask greatly aids the separation of signal from noise.

The process of computing the description may therefore be reduced to three operations: firstly, find the peaks in the measurements obtained from the convolutions of the image with different sizes of mask, and select the relevant peaks using the criterion illustrated in figure 2; secondly, separate the peaks into isolated groups; and thirdly, parse the local configuration of peaks into a descriptive element. A small number of classes of peak configuration suffices to cover the cases that can actually occur, and they are illustrated in figure 3. The figure shows typical combinations of peak patterns that occur in the outputs from edge-mask (upper records) and from bar-mask (lower records) convolutions. Examples of the masks that we use appear in figure 3a. The descriptor EDGE is used when two peaks of about equal and opposite signs occur together in the bar-mask record (3b). If one bar-mask peak is considerably smaller than the other, the edge is classified as an EXTENDED-EDGE (3c). Extended-edges are common where a convex boundary is illuminated from one side. Figure 3d shows an intensity gradient edge, and figure 3e corresponds to the presence of a thin LINE such as can occur in the glare off an object's edge, or a very thin pencil stroke. Finally there are edges that begin and end gradually, and extend over a relatively large distance; these are classified as SHADING-EDGEs (figure 3f). In addition to descriptors of edge type, one can measure an edge's STRENGTH, POSITION, ORIENTATION, and FUZZINESS. This last parameter is computed by comparing the amplitudes of the peaks obtained using masks of the same shape but different sizes. (See figure 2, and Marr (1974b) for the details).

Figure 4 gives an example of an intensity distribution that has been described by this process, and the legend explains which mask convolutions were used. One of the assertions has been traced back to the convolution profiles, and the arrows point to the peaks that gave rise to that particular assertion. The low-level vocabulary that is used here is not intended to be definitive, but some claim is made to the effect that it is a good example of the genre, because it has sufficient expressive power to describe most kinds of shading adequately, and the method is simple and works reasonably well. Experiments are being planned to determine whether the types of intensity change that are distinguished by these primitives are also perceptually distinct.

FIGURE 3



3. Examples of edge- and bar-masks appear in 3a. 3b - f give the classification that is described in the text of peak patterns in edge- and bar-mask convolution profiles. The primary visual processor uses these stereotypes to classify intensity changes in an image.

FIGURE 4

4. The intensity distribution exhibited in 4a, whose profile appears in 4b, was obtained by illuminating a curved piece of white paper from one end, and viewing it from above. Its description, computed using an edge-mask of panel-width 8 (4c), and bar-masks of panel-widths 4 (4d) and 8 (4e), is as follows:

EDGE (POSITION 180) (AMOUNT 136) (FUZZ SHARP)
EDGE (POSITION 312) (AMOUNT 3) (FUZZ 4)
EDGE (POSITION 392) (AMOUNT 2) (FUZZ SHARP)
EDGE (POSITION 535) (AMOUNT -3) (FUZZ 4)
EDGE (POSITION 544) (AMOUNT 25) (FUZZ 5)
EDGE (POSITION 564) (AMOUNT 2) (FUZZ 4)
EDGE (POSITION 590) (AMOUNT 1) (FUZZ 4)
EXTENDED-EDGE (POSITION 682) (AMOUNT -12) (FUZZ 9)
(the peaks giving rise to this edge are marked with arrows)
EDGE (POSITION 724) (AMOUNT -20) (FUZZ 6)
EDGE (POSITION 776) (AMOUNT 3) (FUZZ 4)
EDGE (POSITION 784) (AMOUNT -4) (FUZZ 4)
SHADING-EDGE (POSITION 670) (AMOUNT -14) (WIDTH 67)
SHADING-EDGE (POSITION 491) (AMOUNT 4) (WIDTH 36)
SHADING-EDGE (POSITION 439) (AMOUNT -8) (WIDTH 73)

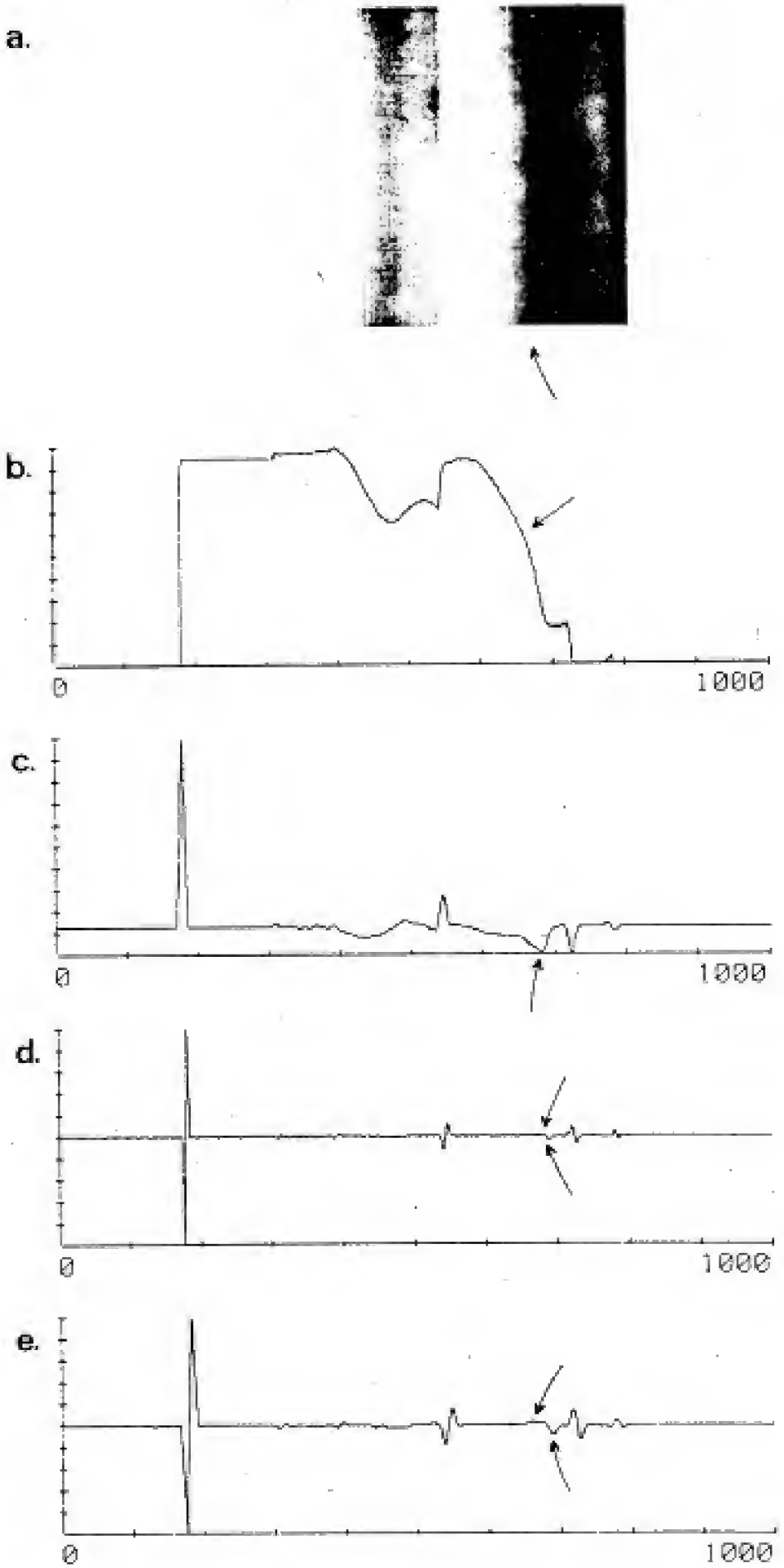
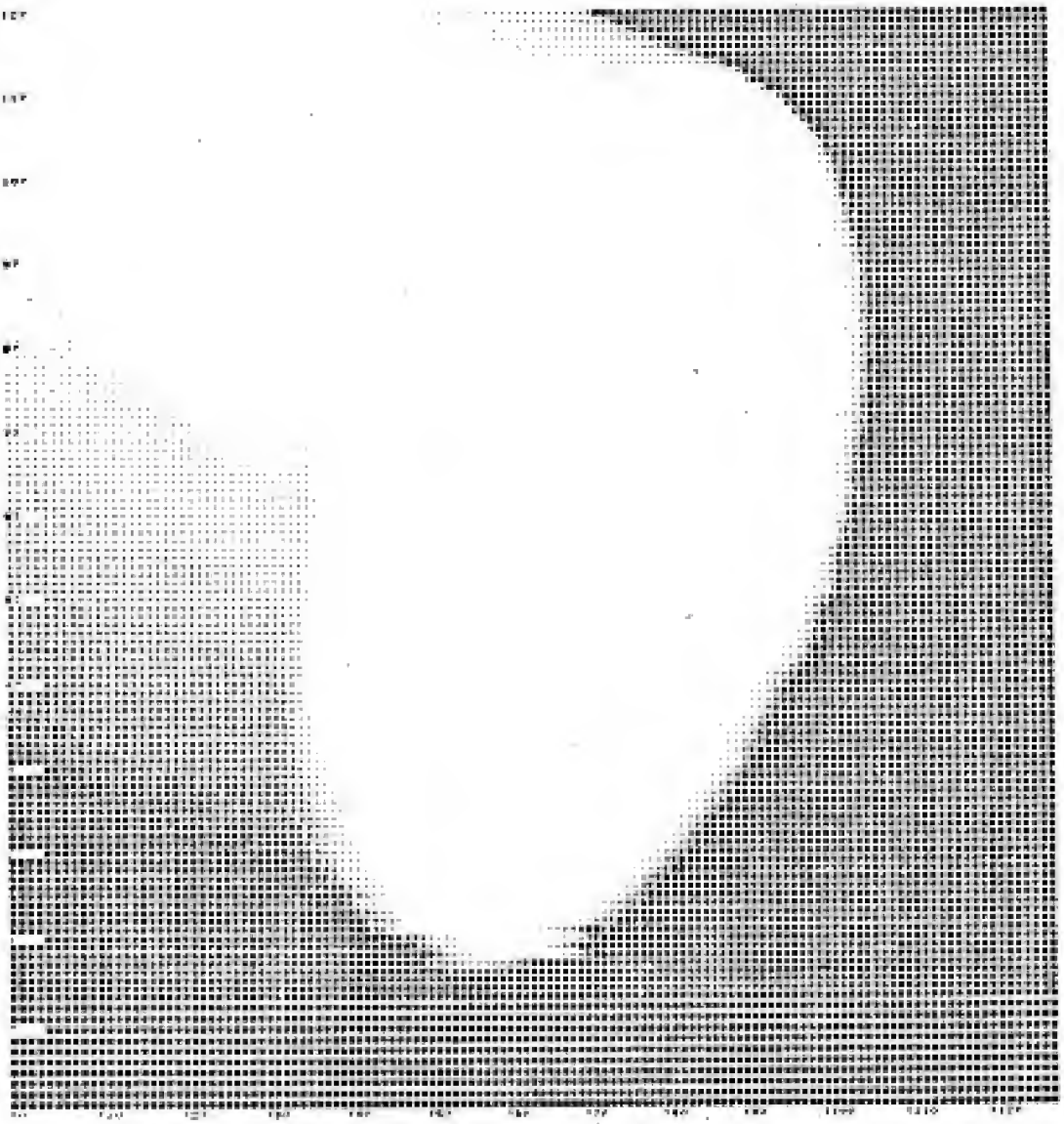

a.
b.
0
1000
c.
0
1000
d.
0
1000
e.
0
1000

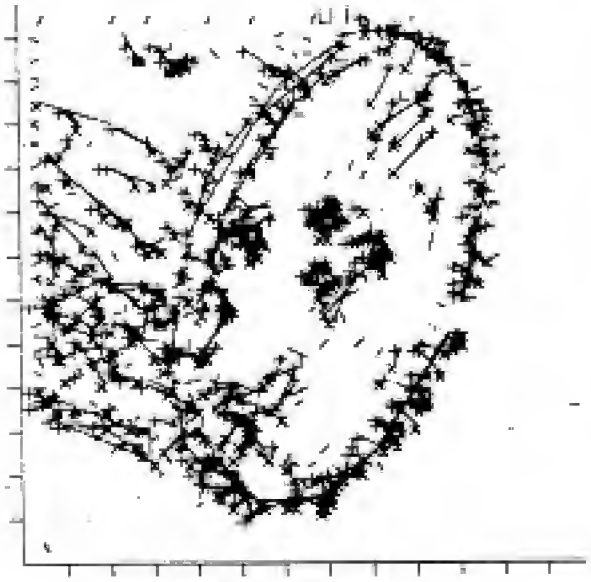
FIGURE 5

5. After description of intensity changes has occurred independently at each of 8 orientations, and after linear assembly of these descriptions has taken place locally, the eight descriptions are combined. An example of the result obtained from 5a appears in 5b. Short noise elimination then takes place, giving 5c. The asterisks denote places at which directional measures of contrast suddenly change. They are the precursors of termination assertions.

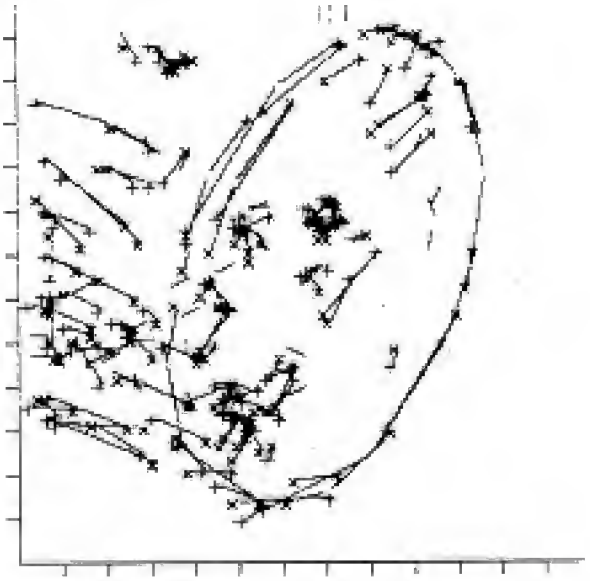
a.



b.



c.



Combining orientation-dependent descriptions

We have seen how to compute an orientation-dependent description of the intensity changes, and we now deal with the problems of combining local pieces of description from the same orientation, and of combining the descriptions obtained at different orientations. What then are the issues that are raised in combining the local analyses described in the previous section?

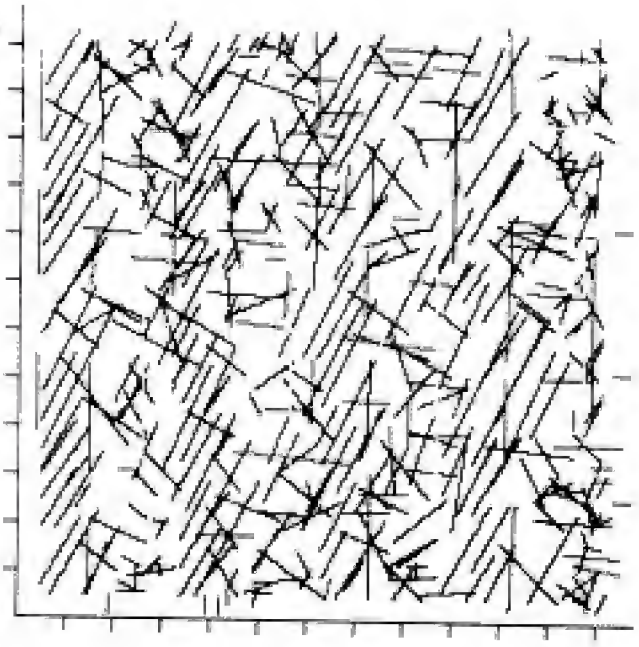
The information that is used during this operation is primarily of two kinds: local consistency relations, which enable one to string local assertions together; and local competition, between competing descriptions of the same phenomenon obtained from masks at different orientations. Surprisingly, it turns out that the local consistency relations are more important than local competition, and that local competition is required not so much between descriptions obtained from masks at nearly adjacent orientations, but between the descriptions obtained from masks that are nearly perpendicular.

Figure 5 illustrates the problems that arise. The image was first operated on at eight orientations with the process described in the last section. Next, these local assertions have been glued along directions nearly parallel to the masks from which they were obtained. An interesting feature of the process is the abundance of short segments perpendicular to the primary edge (figure 5b). These arise because of a combination of local noise, the image tesselation, and other irregularities in the image. They occur in every image we have processed. In dealing with them, one cannot dismiss in a cavalier manner all very short segments: tiny "blobs" in the image also give rise to them, as can be seen from the same image at coordinate (73, 75). But a "small" element like this can be ignored if (a) it crosses a "long" element, and (b) its contrast is less than that of the item it crosses. Figure 5c shows the results of removing small noise elements using this criterion.

The asterisks in the figure signify that the contrast of the edge changes rapidly at that point, possibly becoming zero. They are the precursor of assertions about the presence of terminations, but space forbids a discussion of them here (see Marr 1974c).

The only other item of note in computing the primal sketch is the question of detecting local, small blobs. Figure 5c at coordinate (73, 75) shows how they appear, and in fact we make small blobs a primitive element of the primal sketch, together with their associated "intensity" value, and the sizes and orientations of their major and minor axes. Finding these blobs from the glued assertions depends a small amount on elegant programming, and a large amount on brute force. The reader may ask why do we detect blobs in this way; why not use a simple blob-detector like a mask with a centre-surround organization? The reasons are twofold. Firstly, when using a centre-surround mask to generate assertions, one has to be very careful of the boundary condition problem

b.



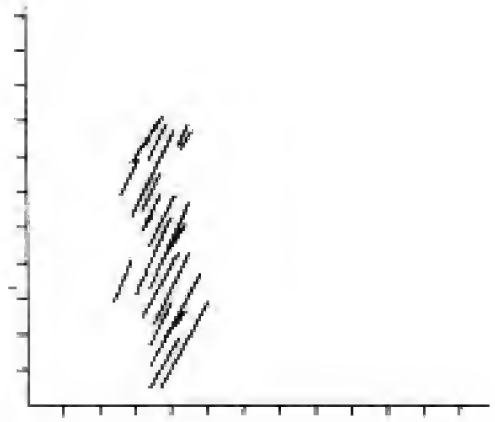
c. COARSE IMAGE DESCRIPTORS

(used in primary control of texture analysis)

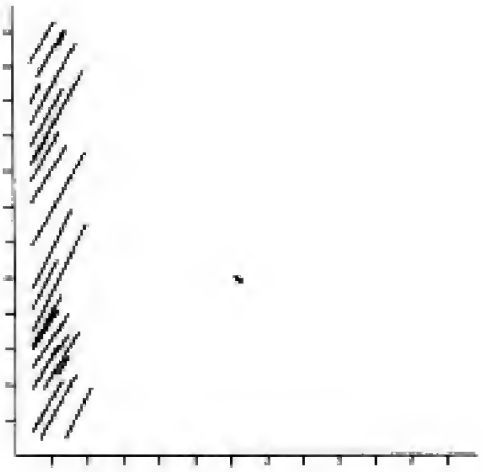
Orientation Brackets are 15° wide

ORIENTATION (degrees)	0	15	30	45	60	75	90	105	120	135	150	165
NUMBER OF ITEMS	64	7	14	16	161	27	42	15	25	28	34	16
TOTAL CONTOUR LENGTH	632	64	132	116	2213	186	600	118	198	304	331	138

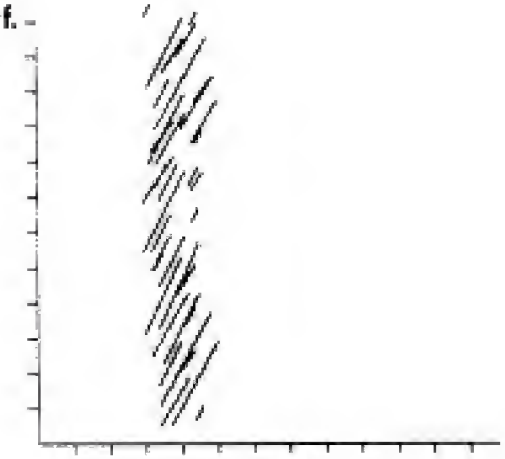
d.



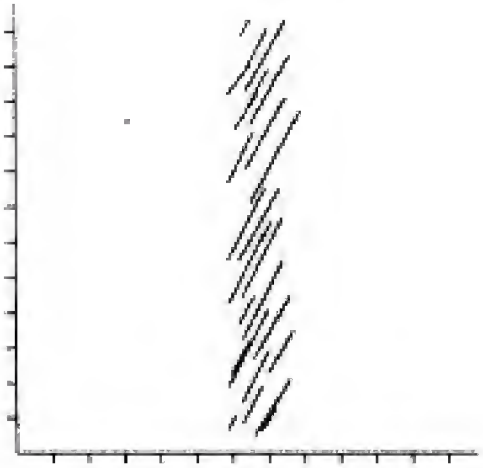
e.



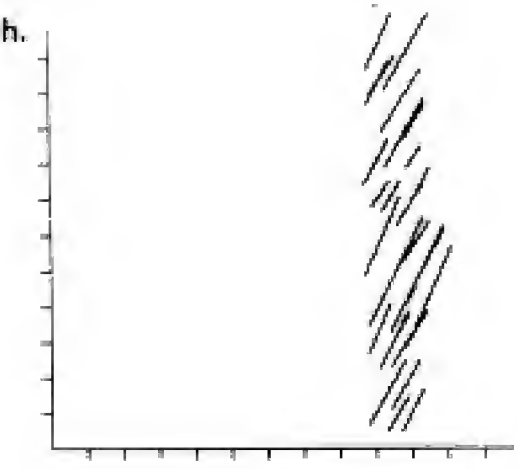
f.



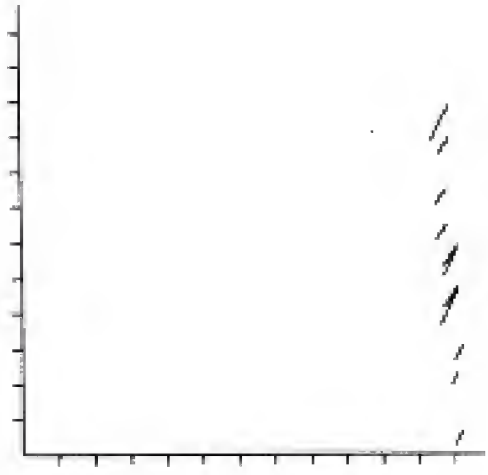
g.



h.



i.



mentioned earlier. One can devise parallel schemes of the form "a blob exists at points P if the centre-surround mask gives an isolated peak there, and if there are no edges in the vicinity," but these are relatively expensive to compute, and become unreliable if the blob is not very circular, or if there are indeed other, fainter or unrelated edges in the vicinity. It is interesting in this connection to note that the phosphenes produced by stimulating a point in area 17 -- an act which presumably stimulates orientation-sensitive cells at all orientations -- commonly take the form of a bright point in the visual field (Brindley 1970, p 124).

The primal sketch differs from a simple feature-point array in a rather subtle way, and as a model of the information-processing that is performed in area 17 it makes some definite and perhaps unexpected statements. Some examples will help to make this clear. One consequence is that the direct output of a linear simple cell is not available as an element in the primal sketch. Its measurement is used to create an assertion about the presence of an edge, and that assertion is what is available. Creating the assertion is an act of computation -- a simple one, since it involves little more than peak matching and the classification of a peak configuration, but an act of computation nonetheless. The main point is that this has to go on.

An interesting consequence of this is illustrated in figure 6. Suppose that an image contains two small close blobs. These blobs give rise to measurements by a number of sizes of mask -- some small ones represented by the tiny line segments, and some large ones, like the one that is illustrated. One's a priori inclination would be to believe that large "line-detector" would fire, and that this would have something to do with seeing the two blobs. This view amounts to supposing that simple cells write directly into a feature-point array. But if our theory is correct, although the large "simple cell" may indeed fire, its measurement will not be used to compute the description of the two blobs because their sharp boundaries cause the associated intensity change to be described from peaks in the small masks. The effect illustrated in figure 2c will cause the description to be computed from the smaller masks unless the blobs are severely defocussed. [Compare also our failure to perceive L. D. Harmon's coarsely sampled and quantized image of Abraham Lincoln, (Julesz 1971, p.311)]. I mention this point because Julesz (1975, pp40-42) has concluded that in situations like this one, the output of large simple cells in this configuration plays no part in texture vision discriminations. We shall see the relevance of this shortly.

The structure of the primal sketch may be summarized as follows:

PS1. The primary visual processor delivers a symbolic description of the intensity changes present in an image. This description uses the following primitives to describe intensity changes:

(i) Various types of EDGE

(ii) LINEs, or thin BARs.
(iii) BLOBs

The items (i) and (ii) have been assembled into straight segments, and short noise elimination has occurred.

PS2. The following items are bound to each element of the description.

(i) ORIENTATION
(ii) SIZE - length and width if both are defined, diameter if major and minor axes are equal or undefined.
(iii) INTENSITY (LIGHTNESS).
(iv) POSITION.
(v) TERMINATION POINTS.

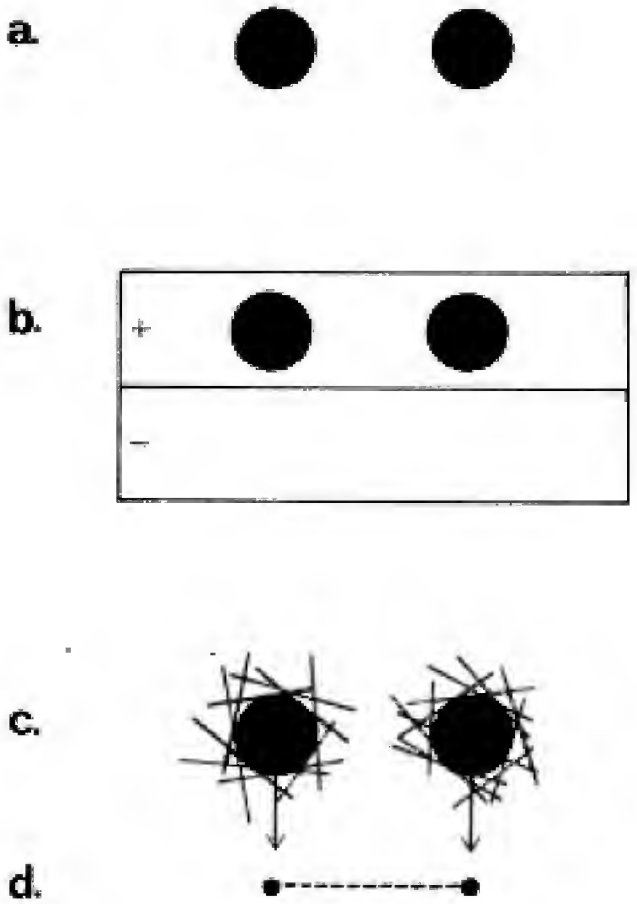
What drawings tell us

In order to make the second step of my argument, I must digress awhile on the manifest variety of ways in which we can interpret simple pencil drawings that lack semantic content. The point I wish to make is that from our ability to interpret certain kinds of drawings, we can infer with some confidence that certain kinds of symbolic process must exist in our visual systems. Let us take an extreme example first. In figure 7a there is little doubt that some process somewhere is creating a circular contour, and that the "places" in the image that are giving rise to that contour are the inner ends of the radial lines. One cannot argue that Fourier detection methods will produce it for one, because it really is not there. This contour is not being detected, it is being constructed. Figure 7b shows another example in which "ends of things" are being formed into a perceptually vivid contour.

From these two rather strong examples, we see that abstractly defined places in an image can be assembled into contours that have a definite perceptual existence, despite the absence of apparent semantic content in the image. If one approaches these phenomena from a computational point of view, it is natural to think of this process as occurring in two steps. Firstly, certain things in drawings can cause "places" to be defined in some abstract sense. Secondly, "places", once defined, can be aggregated in various ways.

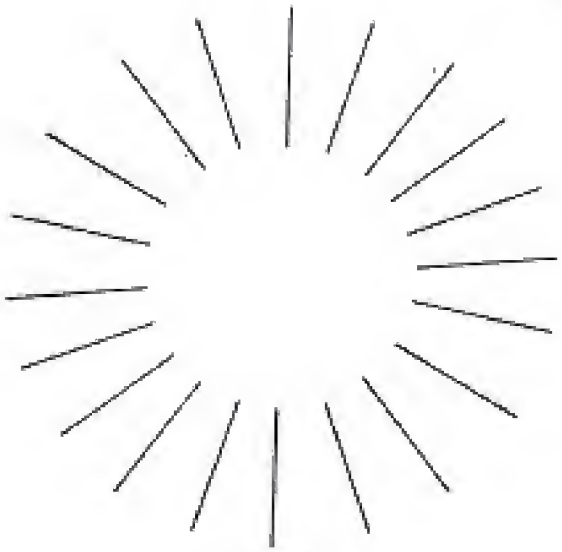
Having realized this, one immediately wants to know in what ways places actually can be defined, and in how many different ways they can be aggregated. A better feel for the problem can be gained by looking at the rest of figure 7, and at figure 8. We are forced to conclude that "places" may carry intrinsic orientation information, and that this orientation information may or may not be used (figures 8d and 7c). Indeed these two situations can occur in the same figure (7e).

FIGURE 6

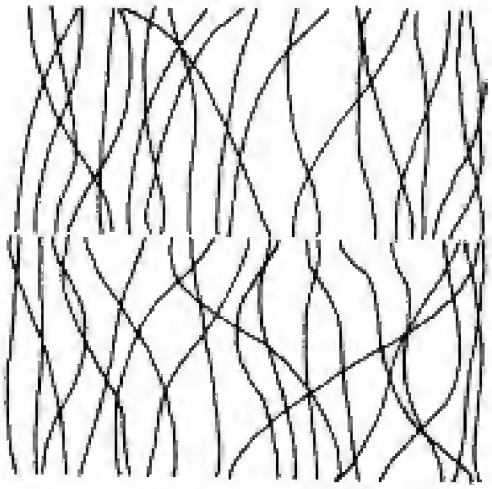


6. The difference between the primal sketch and a feature-point is brought out by the image 6a. A measurement taken with a large mask (6b) could generate a feature-point, but it would not be used in the computation of the primal sketch. This is because the sharp contrast changes force the use of measurements from small masks (6c).

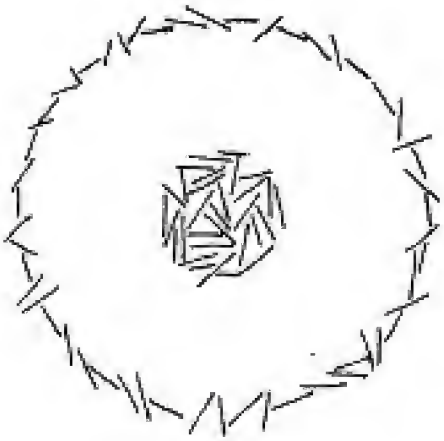
FIGURE 7



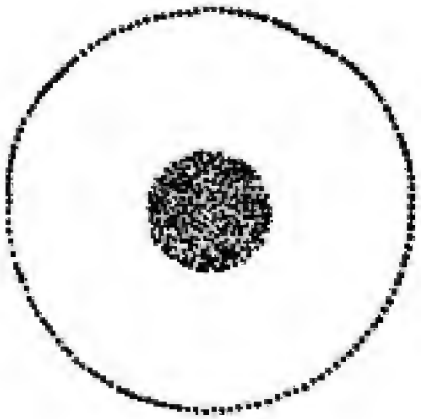
a.



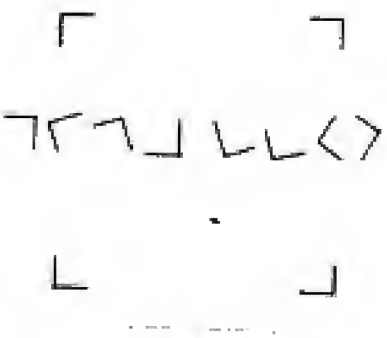
b.



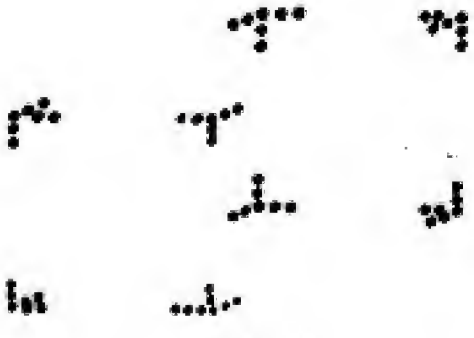
c.



d.



e.



f.

7. These drawings provide evidence for the action of several symbolic processes during our perception of them. In particular, the circular "contour" in 7a, and the linear one in 7b, are being constructed, not detected.

We see from these examples that the aggregation of places can occur in two broad ways: clustering into groups that often have computable boundaries, and the assembling of places into curves or lines, which I call curvilinear aggregation. In the case where there is an orientation associated with the place, aggregation can either use or ignore it. If the orientation is used, there are two possible ways: the aggregation can either follow the intrinsic orientation, or it can proceed in a fixed orientation relative to it (figure 8c). If the number of places involved is very small (less than 5 say), the places may form a standard, named configuration (see figure 9) which is evidently described relative to an axis which is imposed on the figure, and whose default value is the vertical.

Interestingly, procedures for implementing each aggregation technique are quite straightforward. They have a common flavor; a mixture of a simple local process operating everywhere over the image, together with a sensitivity to, and the ability to generate, one or two straightforward global measures. To give you an idea of their simplicity, I shall outline one of them, which we call theta-aggregation. Theta-aggregation is the process by which oriented items are aggregated in a direction that differs from their intrinsic orientation. The difficult part about it arises because measures of the "overlap" of two oriented items depends upon the angle, theta, that the final aggregate makes with each local unit (see figure 10). So theta determines the aggregation process, but also depends upon it. For good data, it may be quite unnecessary to know theta; place aggregation that ignores theta will suffice to compute the aggregate. In general, however, one will need to take theta into account, as we shall shortly see. Viewed from a very abstract level, this computation may be regarded as a process of solving a large number of rather simple equations. In practice, a network with feed-back will solve it, where the information being fed back is theta. We have implemented an iterative version of this process, and some results are displayed later on.

In summary then, the argument of this section has been that our ability to interpret certain simple drawings shows that we can bring certain highly symbolic processes to bear on the analysis of drawings whose semantic content is small. I summarize the processes that appear to be available below, even though space has not permitted mention of several of them.

PLACES may be defined by:

(P1) The position of a blob, or of an edge or line that is not too long.
(P2) The end of an edge or line that is not too short, or of a blob with long major axis and short minor axis.
(P3) A small aggregation of places.

The definition is slightly recursive. This is to be expected, since the assertions

produced by one aggregation process are presumably written into the same active geometrically organized storage processor as is the primal sketch. The precise boundary between "too long" and "too short" can be left to individual taste, because near it, both definitions will usually lead to the same aggregations. The boundary needs to be in the region of 0.5 to 1 degrees of arc at foveal resolution.

AGGREGATION may proceed in the following ways:

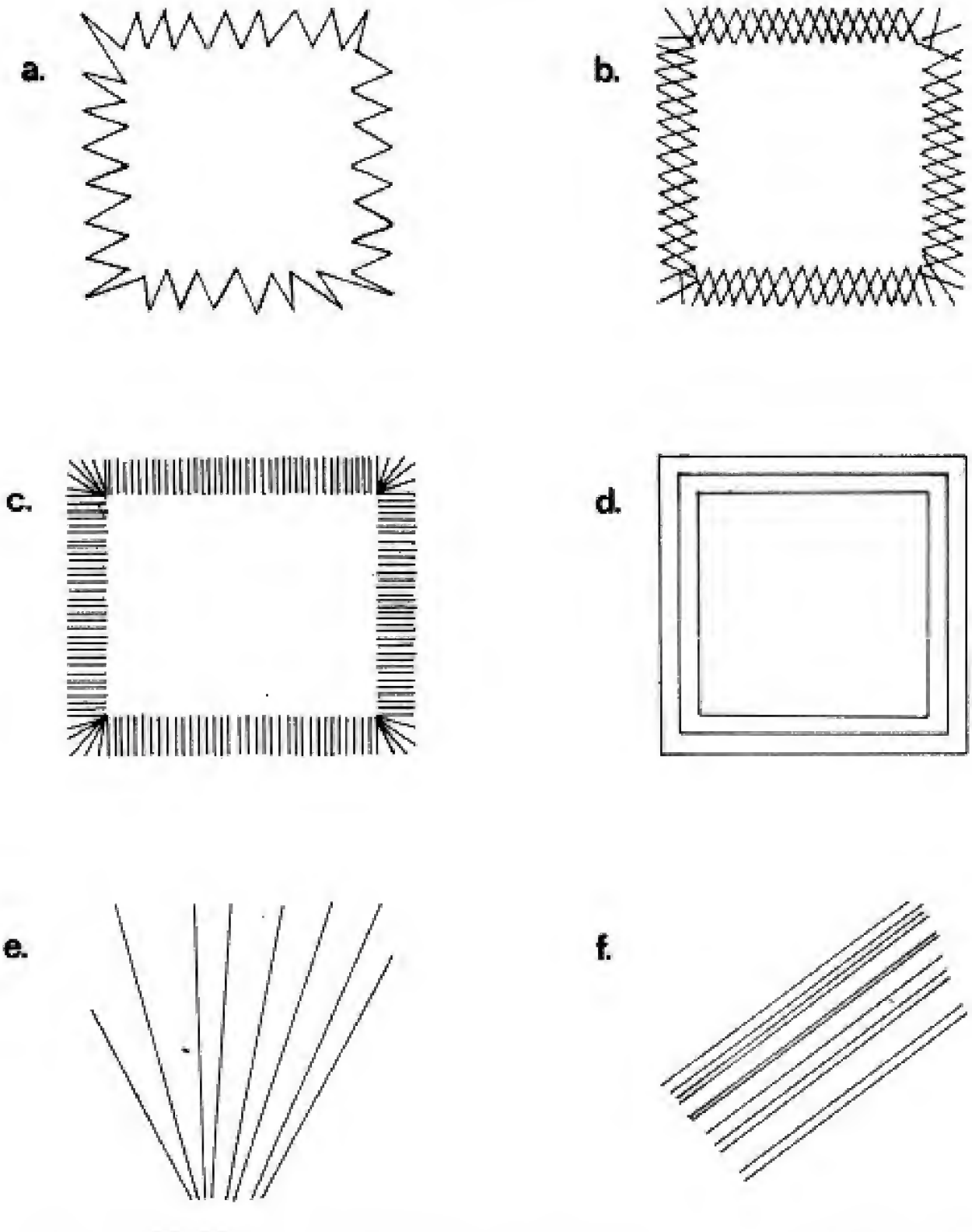
(1) Clustering nearby places, using the methods about as complex as B1 or B2 of Jardine & Sibson (1971), but which are sensitive to global parameters of size and average density. Clustering facilities that appear to have about this complexity can operate on patterns of dots in most human visual systems (see e.g. Julesz (1971 pp 105ff), or recently O'Callaghan (1974a)).
(2) Curvilinear aggregation: aggregation that has a (local) orientation, and which produces contours by joining nearby, aligned places. It is probable that only first and second nearest neighbors need be considered by the local components of these processes, but some global information is also generated and used [see O'Callaghan (1974a and b) for access to recent literature on dot-grouping studies, and Marr (1976)].
(3) Theta-aggregation, the grouping of local, similarly oriented items in a direction that differs from the intrinsic orientation, but in a manner which uses it.
(4) If the number of places is small (< 5), the configuration formed by the places may be described relative to some specified axis by means of a special configuration datastructure (See Marr 1976).

Global Measures on the Primal Sketch

Before the digression of the last section, we had reached the point of defining the Primal Sketch, and of showing how to compute most of the quantities in it. We also saw the primal sketch of a very straightforward image, of a cylinder. The primal sketch is rarely as simple as that, however. Figures 12 and 13 contain examples of the primal sketches of more complex images, and, as one might expect, they are in general large and unwieldy collections of data. Furthermore, it is difficult to see how the complexity of the primal sketch could be an artifact of our particular choice of primitives: images really are complex in this way.

The unwieldy nature of the primal sketch is therefore something with which we have to live, and turn to our advantage if possible. The fundamental problem of the next stage of the analysis is simply stated: how do we select out from the primal sketch those regions that should be treated as unit forms by subsequent descriptive processes; and is it possible to do this without

FIGURE 8



8. These drawings exhibit aggregation processes that take some account of the orientation present at the aggregated places.

complex interactions between the primal sketch and higher-level knowledge? In perceptual terms, the computational problem that we must now address corresponds to distinguishing between figure and ground, and it is strongly related to the problem of texture vision (Julesz 1971).

From an abstract point of view, the primal sketch is simply a large body of data. There is therefore no difficulty in extracting from it certain simple global measures and statistics. In particular, we shall assume that the following measures are automatically available from any primal sketch:

MEASURES taken over moderately sized regions (0.5 to 1.0 degrees at foveal resolution) of the primal sketch:

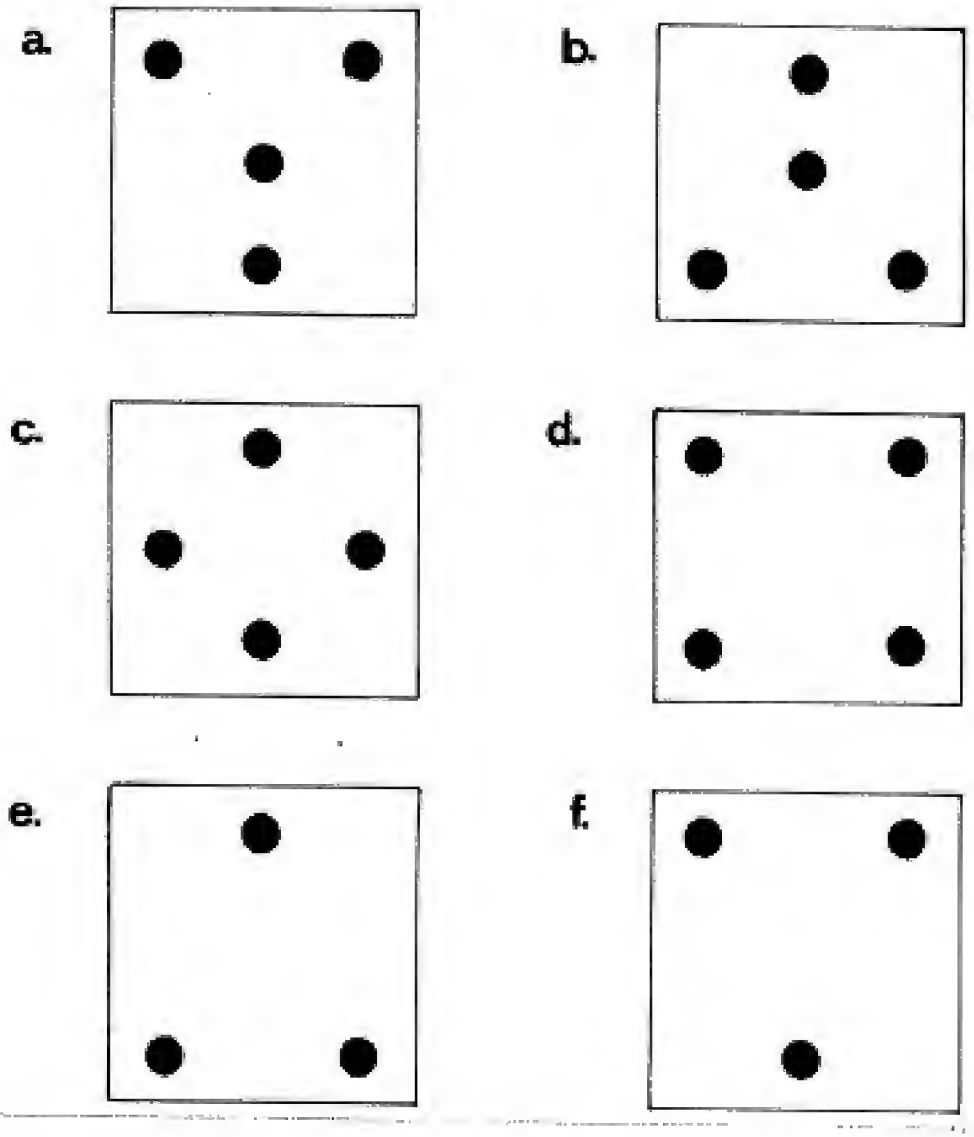
M0. The total amount of contour, and number of blobs, at different contrasts and intensities.
M1. ORIENTATION: the total number of elements at each orientation, and the total contour length at each orientation -- the orientations being divided into about 12 discriminable buckets. Detection of the existence of one, two, or three predominant orientations, and the recognition of distributions that have substantial amounts of contour in more than three orientations.
M2. SIZE: measurement of the mean and variance of the size parameters defined in the primal sketch.
M3. INTENSITY: measurement of the mean and variance of the lightness of items in the primal sketch.
M4. SPATIAL DENSITY: mean and variance of the nearest neighbor distances, and possibly the mean second-nearest neighbor distance. There is no computational problem in obtaining these measures.

Texture Vision

There are three parts to the problem of texture vision. How does one discriminate between textures, and hence form regions from texture differences? How does one describe the shapes and dispositions of the regions so obtained? And finally, how does one interpret a texture, in the sense of understanding the structure of the surface that gave rise to it? Only the first of these will be dealt with here.

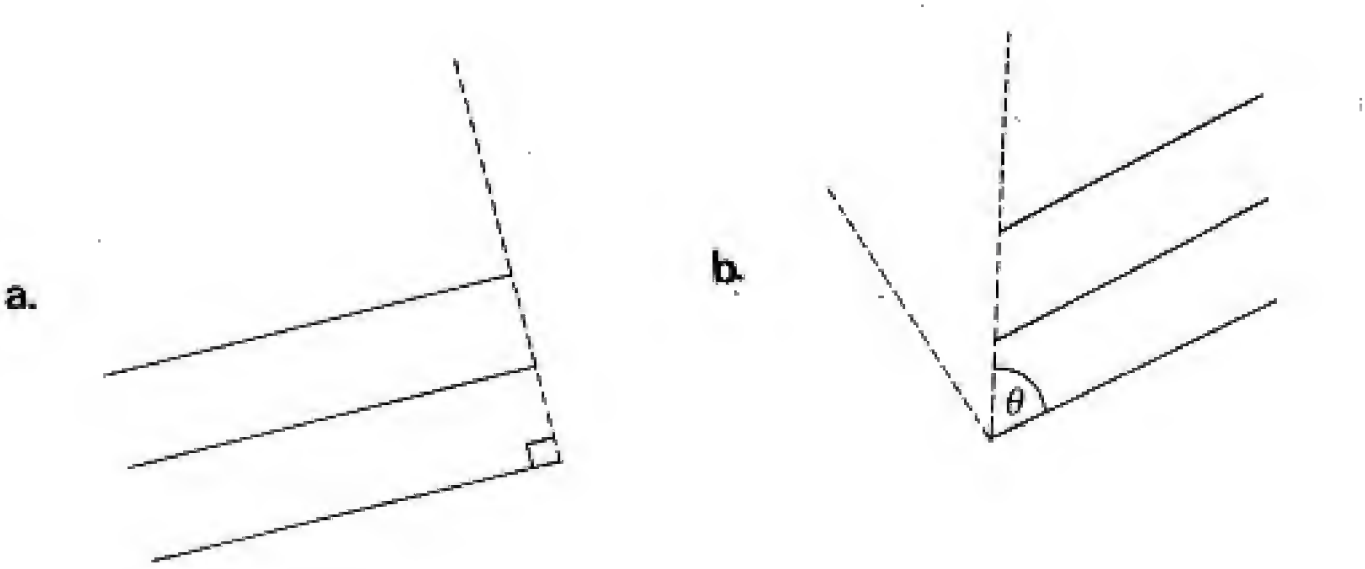
There are several current ideas on texture processing. Some authors have used Fourier techniques, and in certain circumstances, the spatial power spectrum can successfully separate different regions (Bacjsy 1972). Others have constructed specialized operators which when applied to an image sometimes discriminate between regions with different texture. Probably the earliest example of this was the Roberts gradient (Roberts 1963). The most interesting

FIGURE 9



9. Examples of "standard configurations" that we have found it useful to recognise. The reader will probably perceive them relative to a vertical axis. The VEE shown in 9f is used in figure 15e.

FIGURE 10



10. The measure of the overlap of two adjacent, parallel lines depends on an external angle, theta. In 10a, theta is 90 degrees, which is the value at which iteration begins.

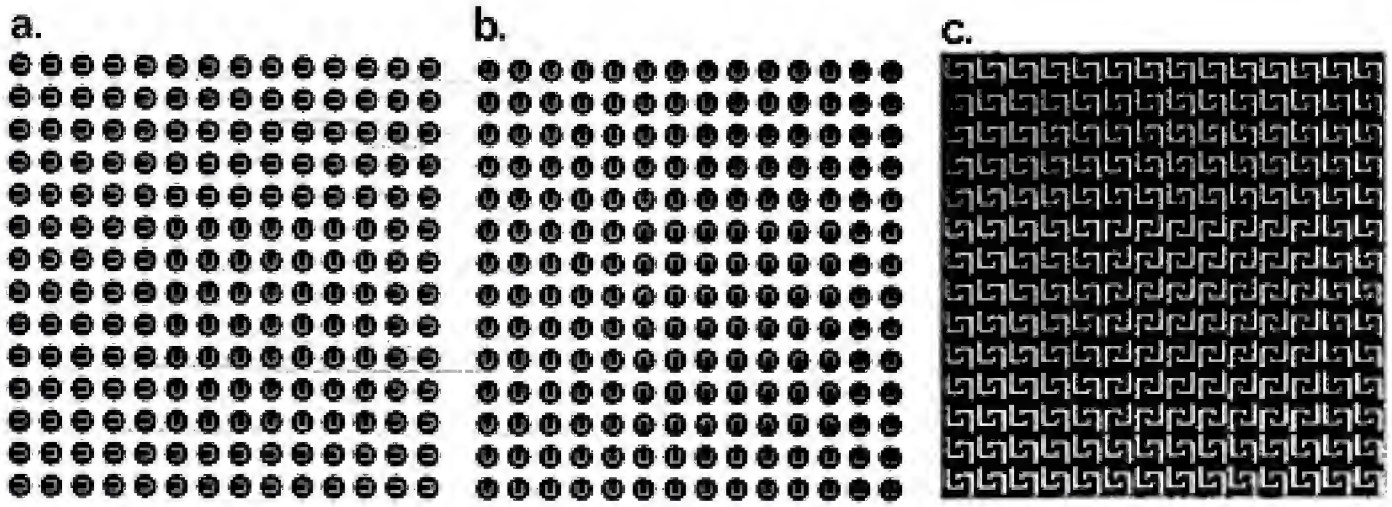
and comprehensive proposal is due to Julesz, Frisch, Gilbert and Shepp (1973), [see also Julesz (1975)], who showed that visual textures that differ only in their third or higher order statistical structure are rarely perceptually discriminable; whereas visual textures that differ in their first or second order statistics can almost always be distinguished. The important point about this finding lies in its demonstration of the essential simplicity of texture processing. Although it gives no insight into the exact nature of the processing, it does imply that all coefficients of third and higher-order terms in its Volterra series expansion are zero.

We have now reached the core of this article. We saw in the last section that certain computational facilities exist and are deployed during our reading of certain kinds of drawings. The facilities were summarized as processes P1-3 and A1-4 on page 14. It is, of course, possible that their existence is no more than a happy accident, which fortuitously allows us to interpret the idle scribblings of the artistically gifted. The central thesis of this article is that these processes are available precisely because they are needed to help interpret the primal sketch; and furthermore that these symbolic processes, together with first-order discriminations based on the measures M0-4 defined on page 15, are sufficient to account for the range of texture discriminations of which we are capable, within the class of images to which this article is restricted. In other words, texture vision is actually implemented not by second-order operations on the image, but by first order discriminations, together with a small number of grouping operations, acting on the primal sketch of the image. Julesz (1975 p43) mentioned in an aside the possibility that texture vision may rest on "first-order statistics of various simple feature extractors", but this idea requires the concepts of the primal sketch and of the aggregation primitives before it can be brought to fruition.

So that the reader may form an intuitive grasp of the central thesis, let us re-examine two of the textures devised by Julesz, and follow this with some examples of the texture analysis run on the images whose primal sketches we saw earlier. Firstly, consider figure 11. Julesz notes that in 11a, the two regions have distinct second-order statistics, but not in figure 11b. Hence, according to his rule, the two regions are distinguishable in 11a, but not in 11b. Now consider our new explanation of this. Orientation measures are the only distinguishing feature of the primal sketch representation, because everything else has carefully been held constant. In 11b, the two basic elements are related by a 180 degree rotation, and so the orientation statistics to which they give rise are identical. Hence the two regions are indistinguishable. In 11a however, there is more contour at 0 degrees than at 90 degrees in the central patch, but the opposite is true in the surround. Hence the two regions are immediately distinguished.

The second example appears as figure 11c. Some of the modules in the pattern have been reflected about a vertical line through their centers.

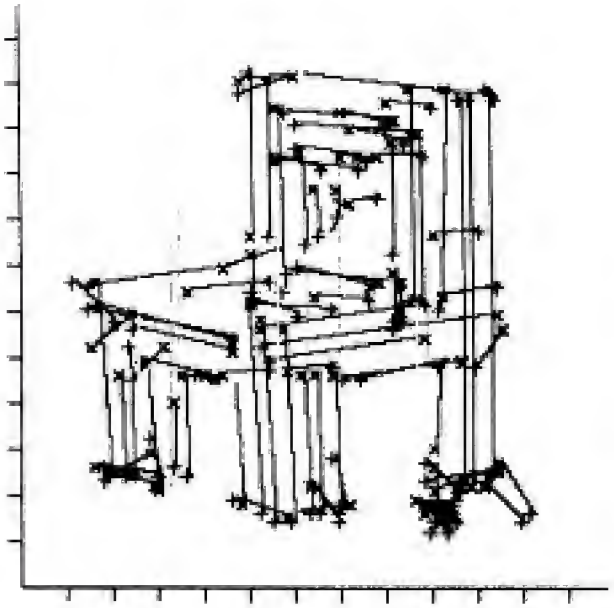
FIGURE 11



11. Examples of textures devised by Julesz. All three contain a square region which differs from the background, but only in 11a is it immediately discernable. The theory provides an explanation of all of them.

FIGURE 12

a.



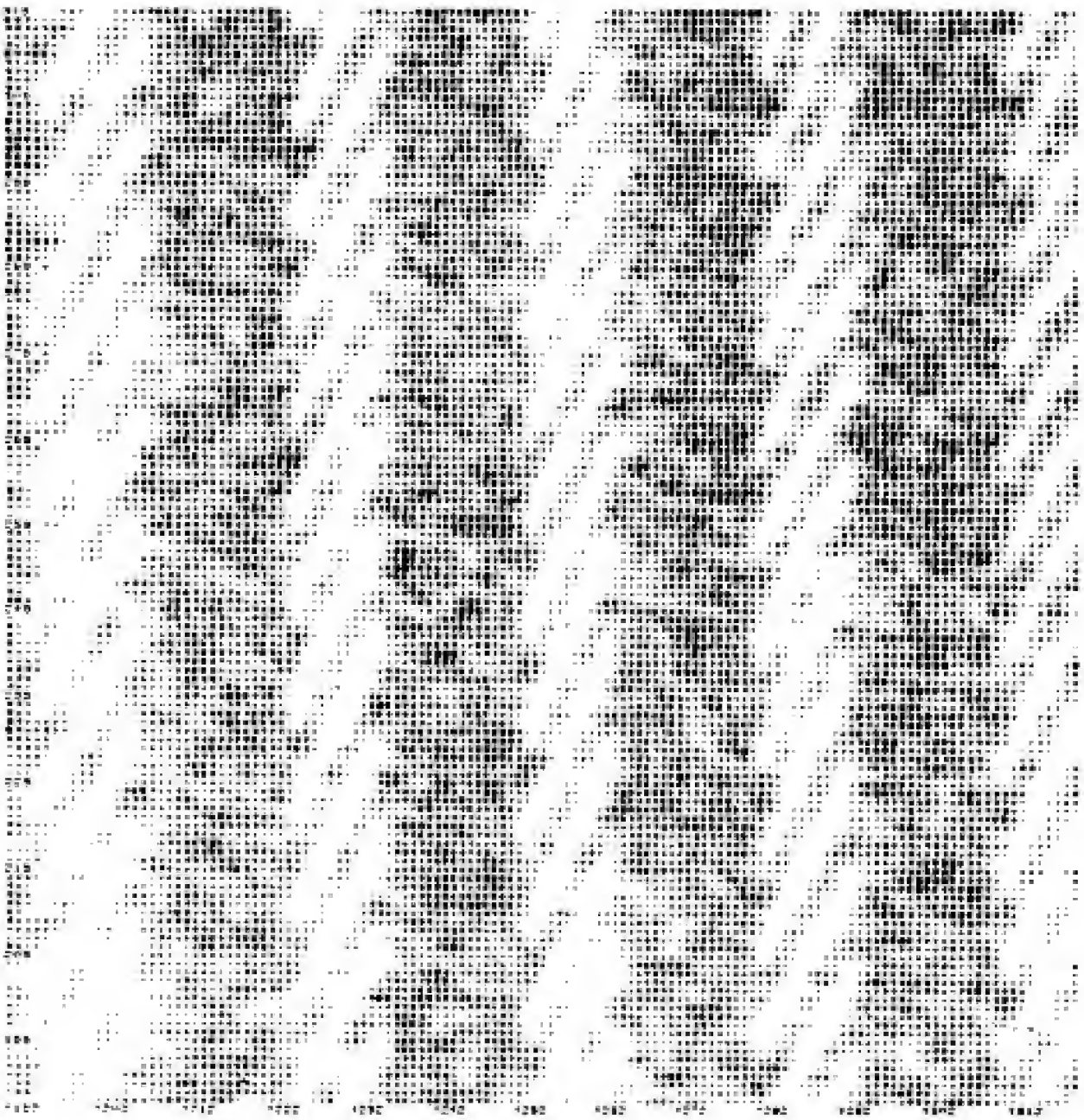
b. COARSE IMAGE DESCRIPTORS
(used in primary control of texture analysis)
Orientation Buckets are 15° wide

ORIENTATION (degrees)	0	15	30	45	60	75	90	105	120	135	150	165
NUMBER OF ITEMS	16	12	2	3	2	1	30	4	2	5	4	14
TOTAL CONTOUR LENGTH	258	264	15	25	14	10	998	34	23	46	25	207

12. 12a gives a rendering of the primal sketch of the image of figure 1a. 12b shows some measures made on it. Theta aggregation has decoded the texture that is present, and the aggregates are displayed as the mosaic 12c.

FIGURE 13

a.



13. 13b shows a rendering of the primal sketch of 13a. 13c gives the associated orientation-dependent statistics. The predominance of items at 60 degrees causes theta-aggregation to be attempted at this orientation. The default setting of theta produces the aggregate 13d. From this, theta is found, and the aggregation process then extracts the stripes successfully (13e - i).

Their second-order statistics are therefore different. This is an example in which Julesz's generalization fails.

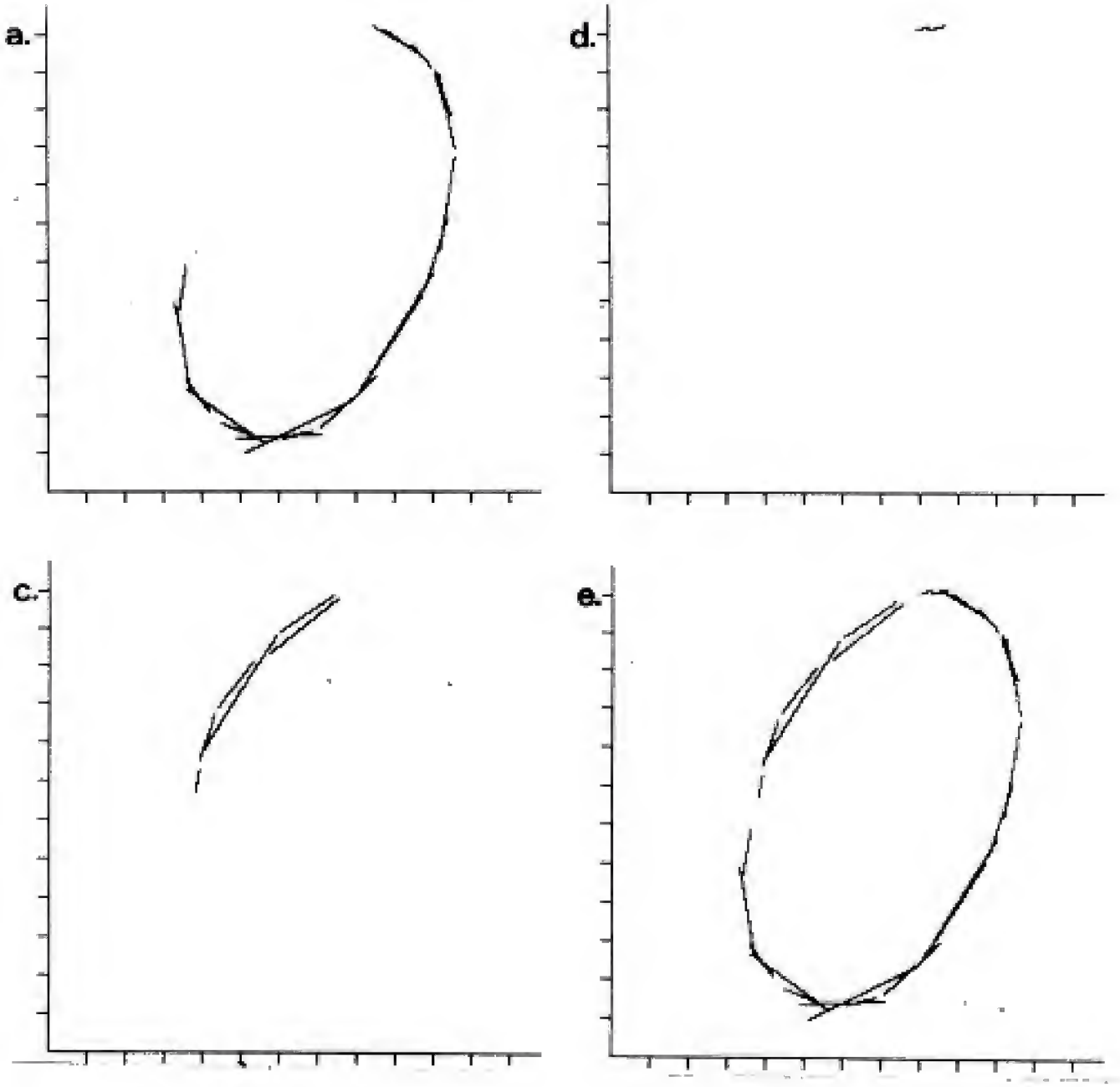
The statistics of the orientations of the contours are however unchanged in this particular instance, because only vertical and horizontal orientations are involved. Hence the present theory predicts that the two regions are in fact indistinguishable.

Now let us look at some real images. Figure 12a shows the primal sketch of the chair whose image appeared as figure 1a, and figure 12b gives some of its orientation statistics. The first thing to realize about this image is that it is textured at all. The texture is so simple that one easily overlooks it. Yet the texture exists in exactly the sense of this article, and the process that succeeds in decoding it is theta-aggregation. Figure 12c shows the results of running the theta-aggregation procedures on this image, and each element in the mosaic contains just one aggregate.

We see from this example a glimmer of the power of texture vision. Using one knowledge-free technique, we have separated the chair from its background, and also separated the problem of divining the overall three-dimensional shape of the chair from the analysis of its surface properties. Each of the aggregates can be described simply by position, orientation, and extent; and this produces a skeleton of the outline of the chair. By considering separately the structure of just one aggregate, one could go on to compute a description of the surface structure of the material out of which the chair is made.

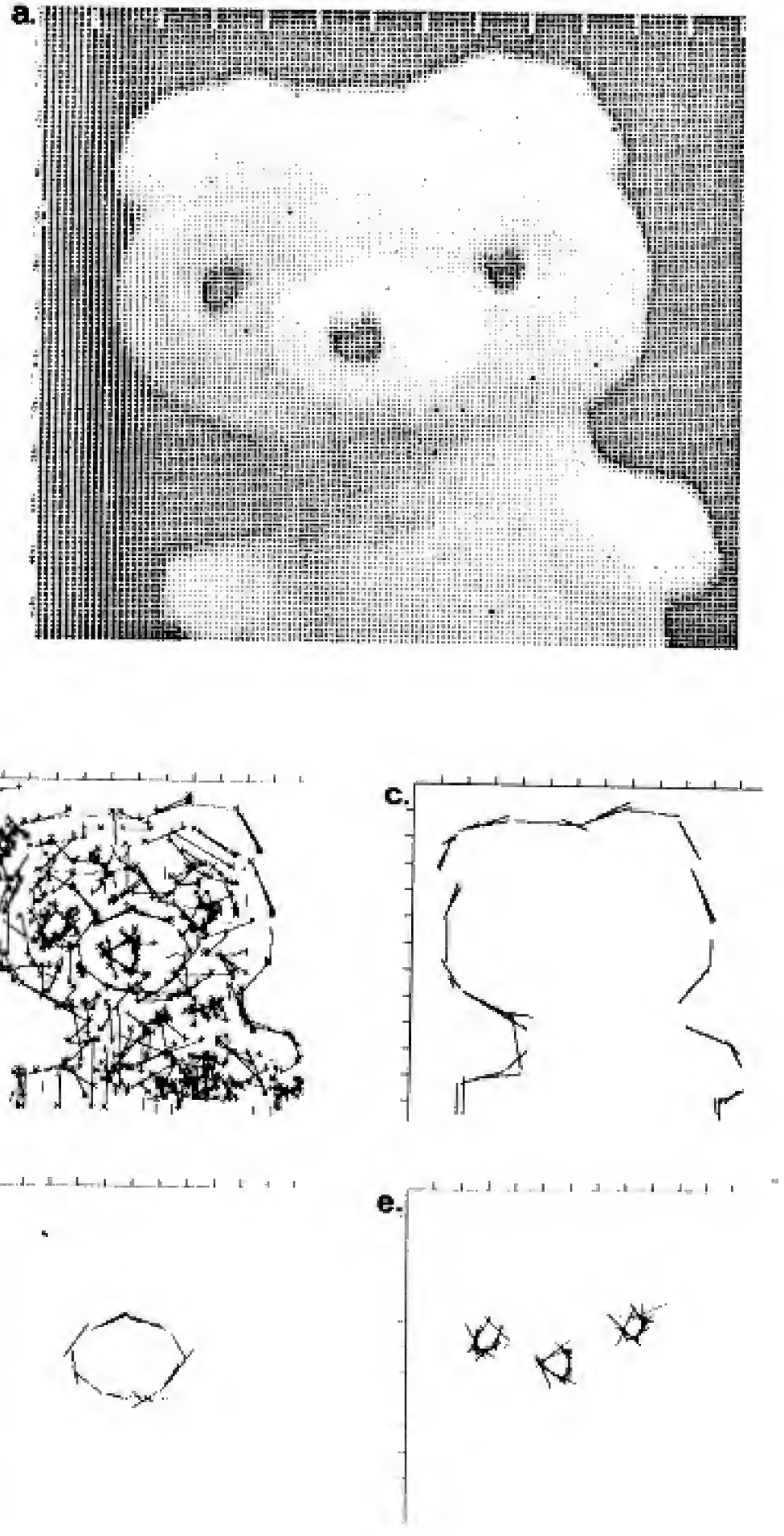
The next example shows a more difficult case of theta-aggregation. The image is taken from Brodatz (1972, plate D11), and the intensity values are shown in 13a. Figure 13b shows an approximation to the primal sketch. Contours of all intensities, lengths, and orientations are shown, and as one would expect from an image of this complexity, 13b has a somewhat messy appearance. Figure 13c gives statistical information about this image, from which it is evident that items at an orientation of around 60 degrees are strongly predominant. The average length of items at this orientation is 13. These coarse measures cause the texture analyzer to attempt to group the edges at this orientation. Initially, the direction in which grouping should take place is unknown, so a default of 150 degs (= 60 + 90) is assumed, and stringent grouping parameters are used. This leads to the primary cluster shown in figure 13d. From this, the correct direction is obtained (-88 degs), and the cluster process then groups the items into the stripes shown in 13e, f, g, h, and i. This completes primary texture processing. Once the primary stripes have been obtained, another stage of theta-aggregation serves to relate the stripes to one-another. Notice that in this image, some of the stripe information has been picked up directly from intensity values (see figure 13b). This would not be true of a more herring-bone texture, and the analysis does not depend upon it. Our present system is successful at processing herring-bone textures of similar complexity in which the two types of stripe have the

FIGURE 14



14. Curvilinear aggregation operating on the primal sketch shown in figure 5c produced the elements 14a, b & c. Once larger units have been obtained, the governing parameters can be relaxed, and the elliptical form (14d) is obtained. At this point, the system is unaware of its shape.

15. This image of a toy bear (15a) has the primal sketch illustrated in 15b. The three principal forms extracted from 15b appear in 15c, d & e. The items in 15e are classed as BLOBs, and the configuration that they form is recognised as a VEE (figure 9f) with modifier FLAT. The axis relative to which this description was computed is the vertical (default value).


a.
b.
c.
d.
e.

same average reflectance.

Next, we give an example of a simple kind of curvilinear aggregation. The local elements of the primal sketch of the cylinder shown in figure 5 are grouped using tight, conservative techniques into the units shown in figure 14a, 14b, and 14c. These are then gathered using slightly weaker constraints into the form shown in 14d. Notice that the contrast across the top-left portion of the form has the opposite sign from the contrast elsewhere. Curvilinear aggregation depends on local information about how well two adjacent segments match; and on global information that includes for example whether the complete form is closed. The global measures can affect the local choice of segment in those infrequent cases where no candidate is to be preferred on purely local grounds (see Marr 1976).

Finally, an example of several types of analysis appears in the image of a toy bear (figure 15a). The primal sketch appears in 15b. The contours of his face and muzzle appear in 15c and 15d, and the three blobs that come from buttons that stand for his eyes and nose appear in 15e. The three blobs define three places, which in turn provoke a specific configuration description relative to the default axis, which is the vertical.

The examples given here do not prove the central thesis of this article. This will need to be tested by experimenting with considerably more images than the twenty or so with which we have dealt hitherto. But they give us grounds for believing it to be a reasonable theory of the computational mechanisms that underlie texture vision and the separation of figure from ground. A more complete report is in preparation (Marr 1976).

The influence of higher-level knowledge and of purpose on visual information processing

Perhaps the most novel aspect of these ideas is the notion that the primal sketch exists as a distinct and circumscribed symbolic entity, computed autonomously from the image, and operated on by a number of local geometrical processes, semi-local measures, and first-order discriminations. In a computational sense, the primal sketch is a very active structure. The information written into it depends on the image, but lurking active in its fabric lie several highly abstract geometrical and statistical processes. It is the direct analog for the class of images studied here of the Cyclopean retina that Julesz (1971) wrote of for binocular vision. More subjectively, it corresponds very closely to the "image" that one is conscious of. This reflects the computational hypothesis that all subsequent analysis reads the primal sketch, not the data from which it was computed. The primal sketch therefore acts in a genuine sense as the interface at which visual analysis becomes a purely symbolic affair.

If it turns out to be true that texture vision is successfully implemented by approximately the set of processes that has been defined in this article, it will mean that visual "forms" can usually be extracted from the image by using knowledge-free techniques. In other words, the extraction of a visual form can usually precede its description. From this it follows that it is usually easy to compute a coarse description of a form.

It is difficult to overstate the importance of this for determining the structure of subsequent recognition processes. It means that one can see the shape of the forest without first computing detailed descriptions of all the trees; that one can compute the cluster of blobs that forms a distant village independently of deciding that some of those blobs are actually buildings. In the more mundane example of figure 15, one can compute that the overall shape of the top form is roughly ovoidal without first having to segment out and describe separately the bumps that are the bear's ears. Furthermore, it suggests that the role of higher level knowledge in this process is not only very restricted, but is also different in kind from its intervention in programs like Shirai's (1973). It does not affect the line-finding stage (the computation of the primal sketch) at all. Its most usual modus operandi is in choosing which processes are to be used to read the primal sketch -- for example by specifying which texture predicate should be used on the image to select the parts of current interest. It can also apply certain limited kinds of flags to critical segments during their aggregation into forms. The coupling between higher-level knowledge and the form-extraction processes is however much weaker than the coupling between the different form-extraction processes.

It is clearly desirable to have some control over which of the possible forms in a figure should be delivered at a given moment from the primal sketch. For example, in the image BEAR there are three possible major forms; the outline of the head, the muzzle, and the three blobs that represent his eyes and nose. It seems probable that only one of these should be made available at a time, and this in turn raises interesting questions about the order in which it is done, the way in which the three forms and their relative positions are described, and the way in which those descriptions trigger a larger datastructure and are absorbed by it. In living systems, which are powerful enough to operate in real time, the control of the direction of gaze may be rather closely related to the order in which these events take place.

Acknowledgements: This study would not have been possible without the advanced and flexible computing facilities that are available at the Artificial Intelligence Laboratory. I thank Hassan Alam, Gary Dudley, and especially Ken Forbus for programming assistance; and Karen Prendergast for preparing the drawings.

References

Bajcsy, R. (1972). Computer identification of textured visual scenes. (Stanford A.I. Lab. Memo. 180.)

Barlow, H.B. (1953). Summation and inhibition in the frog's retina. J. Physiol. (Lond.), 119, 56-68.

Brindley, G.S. (1970). Physiology of the retina and visual pathway. London: Edward Arnold Ltd.

Brodatz, P. (1966). Textures: a photographic album for artists and designers. New York: Dover Publications.

Freuder, E.C. (1975). A computer system for visual recognition using active knowledge. A.I. Lab. Technical report in preparation.

Hubel, D.H. and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. (Lond.), 160, 106-154.

Jardine, N. and Sibson, R. (1971). Mathematical Taxonomy. New York: Wiley.

Julesz, B. (1971). Foundations of Cyclopean Perception. Chicago: The University of Chicago Press.

Julesz, B. (1975). Experiments in the visual perception of texture. Scientific American 232, 34-43 (April issue).

Julesz, B., Frisch, H.L., Gilbert, E.N. and Shepp, L.A. (1973). Inability of humans to discriminate between visual textures that agree in second-order statistics - revisited. Perception, 2, 391-405.

Lettvin, J.Y., Maturana, H.R., McCulloch, W.S. and Pitts, W.H. (1959). What the frog's eye tells the frog's brain. Proc. Inst. Radio Engrs., 47, 1940-1951.

Maffei, L. and Fiorentini, A. (1973). The visual cortex as a spatial frequency analyser. Vision Res., 13, 1255-1267.

Marr, D. (1974a). On the purpose of low-level vision. (To appear: preliminary version available as M.I.T. A.I. Lab. Memo 324).

Marr, D. (1974b). The low-level symbolic encoding of intensity changes in an image. (To appear: preliminary version available as M.I.T. A.I. Lab. Memo 325).

Marr, D. (1974c). The recognition of sharp, closely spaced edges. (To appear: preliminary version available as M.I.T. A.I. Lab. Memo 326).

Marr, D. (1976). Configurations, regions, and simple texture vision. (In preparation).

Mountcastle, V.B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. J. Neurophysiol., 20, 408-434.

O'Callaghan, J.F. (1974a). Human perception of homogeneous dot patterns. Perception, 3, 33-45.

O'Callaghan, J.F. (1974b). Computing the perceptual boundaries of dot patterns. Computer graphics and image processing, 3, 141-162.

Ratliff, F. (1965). Mach Bands: quantitative studies on neural networks in the retina. San Francisco: Holden-Day.

Roberts, L. (1963). Machine perception of three-dimensional solids. Technical Report 315, Lincoln Laboratory, M.I.T.

Shirai, Y. (1973). A context-sensitive line finder for recognition of polyhedra. Artificial Intelligence, 4, 95-120.

Thomas, A.J. & Binford, T.O. (1974). Information processing analysis of visual perception; a review. (Stanford A.I. Lab. Memo. 227.)